

Controllable Motion Synthesis and Reconstruction with Autoregressive Diffusion Models

Wenjie Yin¹, Ruibo Tu¹, Hang Yin¹, Danica Kragic¹, Hedvig Kjellström¹, Mårten Björkman¹

Abstract—Data-driven and controllable human motion synthesis and prediction are active research areas with various applications in interactive media and social robotics. Challenges remain in these fields for generating diverse motions given past observations and dealing with imperfect poses. This paper introduces MoDiff, an autoregressive probabilistic diffusion model over motion sequences conditioned on control contexts of other modalities. Our model integrates a cross-modal Transformer encoder and a Transformer-based decoder, which are found effective in capturing temporal correlations in motion and control modalities. We also introduce a new data dropout method based on the diffusion forward process to provide richer data representations and robust generation. We demonstrate the superior performance of MoDiff in controllable motion synthesis for locomotion with respect to two baselines and show the benefits of diffusion data dropout for robust synthesis and reconstruction of high-fidelity motion close to recorded data.

I. INTRODUCTION

Motion synthesis techniques play an important role in computer animation, video games, human-robot interaction [1], etc. Recently, significant progress has been achieved in motion generation and reconstruction by utilizing deep generative models [2], which can be broadly divided into deterministic and probabilistic models. Deterministic models [3], [4] frame the motion synthesis task as a regression problem in which the response and input have exact relationships, leading to stereotypical results with limited diversity. In contrast, probabilistic models fit probabilistic distributions to the data distribution [5], [6], which capture a range of motion patterns and as such significantly improve the motion diversity and fidelity.

Despite recent advances in deep generative models, motion synthesis still remains challenging in a number of aspects. For example, capturing complex relations between body limbs and motion frames requires models that are less susceptible to failures such as mode collapse. Also, robust and coherent synthesis is desirable even when long-term generation is conditioned on imperfect data. The latter is particularly demanding when human skeletal data are extracted from noisy sensors or previously generated frames. Earlier work often assume perfect conditioning patterns [5] or manually defined graph structures [2], and thus cannot satisfactorily mitigate this issue.

In this paper, we propose MoDiff, a diffusion-based probabilistic model for high-quality controllable human motion synthesis, as illustrated in Fig. 1. Diffusion-based approaches

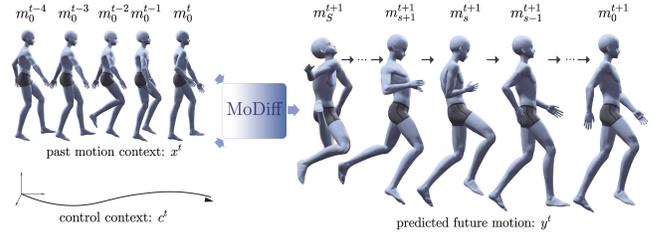


Fig. 1: Illustration of the generation process: MoDiff progressively denoises a noisy motion sequence to get a natural movement, given past motion and control contexts.

have recently gained traction for their superior performance as probabilistic generative models and found application also in human motion synthesis [7], [8]. Score-based diffusion models require no special neural network architectures, such as in the case of flow-based approaches [5], [2], to optimize an exact likelihood. Our work leverages this flexibility by introducing a cross-modal Transformer-based architecture, which enables richer representations for encoding past motion frames and control contexts. The approach also exploits the intermediate representations generated in the diffusion process for a natural dropout strategy to improve robustness. Our evaluation on domain standard human locomotion datasets shows that MoDiff outperforms baseline methods and produces realistic motions conditioning on control contexts of other modalities. We further demonstrate that the same framework can be applied for reconstruction of imperfect motion sequences.

In summary, our contributions are:

- We present a flexible neural architecture, MoDiff, that integrates multimodal transformer and autoregressive diffusion models for motion generation and reconstruction of missing parts in motion sequences.
- We propose diffusion data dropout, utilizing the forward process to obtain diffusion-induced motion representations, which can be employed to improve adherence to the control context.
- The proposed approach achieves superior results for human motion generation and reconstruction. Applications include, but is not limited to, locomotion synthesis, text-to-motion, and music-to-dance.

The paper proceeds as follows: Section II gives an introduction to previous work on human motion synthesis, diffusion models, and data dropout. Section III formulates the problem and provides a comprehensive description of the proposed framework for controllable motion synthesis and

¹Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Sweden. {yinw, ruibo, hyin, dani, hedvig, celle}@kth.se.

reconstruction. Section IV introduces the experimental setup, discusses the results, and compares these to two baselines. Finally, Section V summarizes the paper and outlines future directions.

II. RELATED WORK

In this section, we provide an overview of deep learning-based human motion synthesis (Section II-A) and denoising diffusion probabilistic models (DDPMs) (Section II-B), and then describe prior work on data dropout (Section II-C).

A. Human Motion Synthesis

Deep learning approaches have been widely adopted for human motion synthesis following earlier success in other domains. These approaches can be categorized into deterministic and probabilistic methods. Most previous works follow the deterministic approach to yield a fixed output for a given input. For instance, Fragkiadaki et al. [3] first applied recurrent neural networks to human motion prediction. Butepage et al. [4] directly fed the most recent previous frames through an encoder-decoder network to predict future motion frames. Li et al. [9] introduced a conditioned LSTM for synthesizing long-term motion patterns, and Martinez et al. [10] employed a sequence-to-sequence architecture with residual connections for joint prediction.

To synthesize motion patterns with more variety and diversity, probabilistic methods have also been adopted. Earlier works model distributions over human motion with Gaussian mixture models [11], and Gaussian Process models [12]. As for deep neural networks, Variational Autoencoders (VAEs) and their variants, which optimize a lower bound on the data log-likelihood, have been used in combination with recurrent structures for controllable motion synthesis [13]. VAE-based approaches have also been utilized in cross-modal synthesis tasks, such as generating motion sequences from speech [14]. Another significant branch of methods is based on Generative Adversarial Networks (GAN) [6], [15]. GANs are deemed more powerful and effective for encoding, but are often difficult to train and evaluate. To this end, Flow-based generative models gained in popularity, as they enable tractable likelihood evaluation and efficient model parameterization. MoGlow [5] models motion sequences with autoregressive normalizing flow and recurrent neural networks. Yin et al. [2] further integrated MoGlow with graph convolutional networks for motion reconstruction. Ho et al. [16] presented diffusion models, a new paradigm for probabilistic generative modeling that allows for greater flexibility in the choice of architectures. Our method adopts a diffusion model for improved model capacity and training stability in learning controllable motion synthesis.

B. Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs [16], [17] are a class of generative models inspired by non-equilibrium thermodynamics. DDPMs have a forward and a reverse process. The forward process progressively adds noise to data and the reverse process learns to construct data samples from the noise. The learning model

allows regular neural networks to be used, which makes the models both analytically tractable and flexible.

Several works on human motion generation have adopted diffusion models. MID [18] encodes historical information on behaviors and global signals as an embedding and devises a Transformer-based diffusion model for human trajectory prediction. BelFusion [19] is a diffusion model that exploits a behavioral latent space for human motion prediction. The recent MotionDiffuse [7] is considered to be the first diffusion model-based text-driven motion generation framework with instructions on body parts. Similar to MotionDiffuse, MDM [8] integrates diffusion models and CLIP [20] for text-to-motion generation. Recently, diffusion models have been utilized to generate dance movements, a challenging task due to the intricate postures, rhythms, and compositions involved in dance, often accompanied by music. Alexanderson et al. [21] pioneer diffusion models with Conformer [22] for audio-driven dance motion generation. Tseng et al. [23] propose EDGE, a transformer-based diffusion model that generates dance sequences conditioned on music.

Our proposed framework applies diffusion models for motion generation and models the temporal information in an autoregressive manner, similar to TimeGrad [24], with a cross-modal transformer inspired by Li et al. [25] for controllable generation from various modalities. The difference is that our framework can impose/alter the control signal on the motion generation process on the fly while other models need a supplied command, whether it be text-based or in other forms, before the full sequence can be generated. Moreover, the design of the autoregressive diffusion model is flexible and can be extended to tasks that require robust generation, e.g. synthesis and reconstruction from imperfect motion frames, without additional training, something that is not featured in the works reviewed above.

C. Data Dropout

Due to an over-reliance on the autoregressive context, autoregressive models often suffer from poor adherence to the control context and as such have compromised consistency in controllable generation [5]. Such a phenomenon is exacerbated with long-term prediction. Natural approaches to counter this problem includes removing some or all of the conditioning information during learning. Bowman et al. [26] randomly replace some part of the conditioned word tokens with a generic unknown word token. They apply this technique to a decoder, helping the model to capture higher-order statistics. Wang et al. [27] propose a data dropout strategy that randomly sets the data to zero in both the training and generation stages, forcing the models to focus on the control context, thus alleviating this problem. Kovács et al. [28] propose an input channel dropout scheme, forcing the network to make decisions based on a subset of channels. To remedy the poor adherence issue in the original MoGlow [5], MoGlow applies dropout to entire frames of the autoregressive past motion context.

In this paper we propose a new diffusion-induced dropout scheme. We leverage the Gaussian noise injected in the dif-

Algorithm 1: Training for data sample at time t

Input: data y_0^t , past motion x^t , and control input c^t
Repeat
 Initialize $s \sim \text{Uniform}((1, \dots, S), \epsilon \sim \mathcal{N}(0, \mathbf{I}))$
 if diffusion dropout, $p \sim \mathcal{U}(0, 1)$ **do**
 $x_d^t = x_s^t$ if $p < p_d$, else, $x_d^t = x^t$
 Take gradient step on
 $\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_s}y_0^t + \sqrt{1 - \bar{\alpha}_s}\epsilon, s, x_d^t, c^t)\|$
Until converged

fusion forward process and use intermediate representations as corrupted conditions. Our strategy naturally removes the requirement of tuning a separate process and is found to benefit the consistency between the generated motion and the control contexts. In addition, this also encourages encoders to learn richer and more robust representations.

III. METHODOLOGY

This section formulates our target problem and establishes notations used throughout the paper. Preliminaries about denoising diffusion probabilistic models are also given, including the training and inference strategies. On the basis of these, we introduce our contributed framework.

A. Problem Formulation

In our scenario, we treat the human motion sequence as a series of poses, and the aim is to synthesize the future motion and reconstruct the past imperfect poses using an autoregressive diffusion model. Formally, a 3D skeleton-based pose at time step t is denoted as m^t , with corresponding additional information a^t , such as control signals, texts, music pieces, etc. For the synthesis task, the input of the diffusion framework is the past human poses $x^t = \{m^{t-T_h}, m^{t+1-T_h}, \dots, m^t\}$, and the control input $c^t = \{a^{t-T_h}, a^{t+1-T_h}, \dots, a^{t+T_p}\}$, where T_h denotes the length of the observed past poses and T_p denotes the number of predicted frames. The output of the framework is the predicted future motion frames, written as $y^t = \{m^{t+1}, \dots, m^{t+T_p}\}$. For the reconstruction task, the past human poses are partially observed, e.g. with missing frames or missing body joints. In such cases, the task is to reconstruct a complete motion x^t from an imperfect input, that we denote \hat{x}^t , with the control input c^t .

B. Motion Diffusion Models

We address the formulated problem with our proposed autoregressive diffusion model (MoDiff) based on DDPM [16], as illustrated in Fig. 2. We define the forward diffusion process as (y_0, y_1, \dots, y_S) , where S is the maximum number of diffusion steps. For the sake of brevity, we omit the superscript t . The forward process is a stochastic process with a fixed Markov chain that gradually adds Gaussian noise to the ground truth future motion data $y_0 = y$ until the distribution of y_S is close to a standard Gaussian distribution:

Algorithm 2: Inference the future motions y_0^t

Input: noise $y_S^t \sim \mathcal{N}(0, \mathbf{I})$, past motion x_t , and control input c^t
for $s = S, \dots, 1$, **do**
 $y_{s-1}^t = \frac{1}{\sqrt{\alpha_s}}(y_s^t - \frac{\beta_s}{\sqrt{1-\bar{\alpha}_s}}\epsilon_{\theta}(y_s^t, s, x^t, c^t)) + \sqrt{\beta_s}\mathbf{z}$
 where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ if $s > 1$, else $\mathbf{z} = 0$
end for
Output: y_0^t

$$q(y_{1:S}|y_0) := \prod_{s=1}^S q_{\theta}(y_s|y_{s-1}); \quad (1)$$
$$q(y_s|y_{s-1}) := \mathcal{N}(y_s; \sqrt{1 - \bar{\beta}_s}y_{s-1}, \beta_s \mathbf{I}),$$

where $\beta_1, \beta_2, \dots, \beta_S$ are the fixed variance schedulers for controlling the noise scale. As shown in [16], the forward diffusion sample at any diffusion step s can be calculated in one step as:

$$q(y_s|y_0) := \mathcal{N}(y_s; \sqrt{\bar{\alpha}_s}y_0, (1 - \bar{\alpha}_s)\mathbf{I}), \quad (2)$$

where $\alpha_s = 1 - \beta_s$ and $\bar{\alpha}_s = \prod_{i=1}^s \alpha_i$. In the reverse generation process, we learn this process as $(y_S, y_{S-1}, \dots, y_0)$ and generate motions by progressively denoising the pose from y_S to y_0 . We model this reverse generation process by parameterizing Gaussian transition probabilities with the past poses x and the past and current control signal c as conditioning information. The reverse generation process is formulated as:

$$p_{\theta}(y_{0:S}|x, c) := p(y_S) \prod_{s=1}^S p_{\theta}(y_{s-1}|y_s, x, c)$$
$$p_{\theta}(y_{s-1}|y_s, x, c) := \mathcal{N}(y_{s-1}; \mu_{\theta}(y_s, s, x, c), \Sigma_{\theta}(y_s, s)), \quad (3)$$

where $p(y_S)$ denotes a prior noise Gaussian distribution and $\Sigma_{\theta}(y_s, s) = \beta_s \mathbf{I}$. All transitions share the same parameters.

C. Training and Inference

With the formulated forward diffusion process and reverse generation process, to generate the human pose of future motion frames y_0 , the training process optimizes the log-likelihood in the reverse generation process by maximizing the variational lower bound:

$$\mathbb{E}[\log p_{\theta}(y_0)] \geq \mathbb{E}_q[\log \frac{p_{\theta}(y_{0:S}, x, c)}{q(y_{1:S}|y_0)}]$$
$$= \mathbb{E}_q[\log p(y_S) + \sum_{s=1}^S \log \frac{p_{\theta}(y_{s-1}|y_s, x, c)}{q(y_s|y_{s-1})}]. \quad (4)$$

We utilize the negative bound as the loss function, written as the KL-divergence between Gaussian distributions:

$$\mathbb{E}_q[-\log p_{\theta}(y_0|y_1, x, c) + D_{KL}(q(y_S|y_0)||p(y_S))$$
$$+ \sum_{s=2}^S D_{KL}(q(y_{s-1}|y_s, y_0)||p_{\theta}(y_{s-1}|y_s, x, c))]. \quad (5)$$

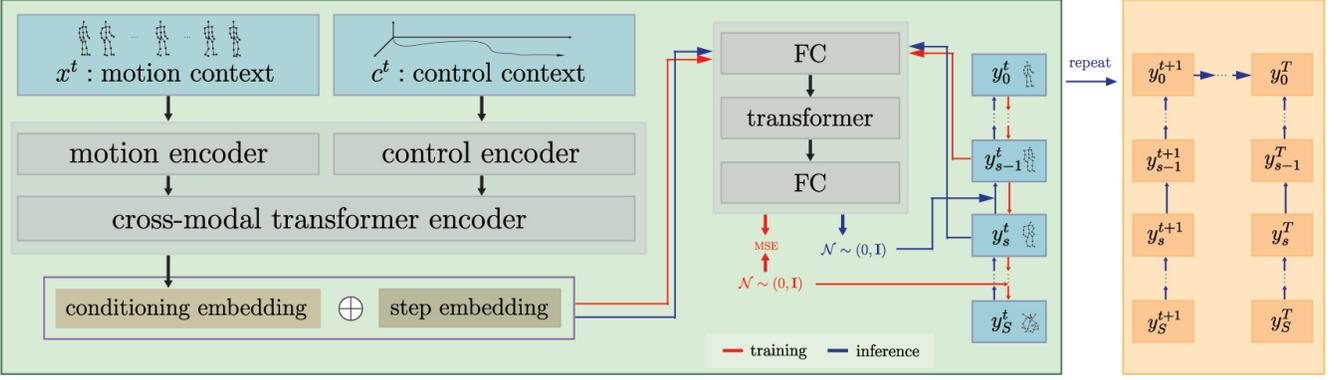


Fig. 2: Overview of the MoDiff schematic: a Transformer-based diffusion probabilistic model synthesizes future motion by the learned reverse diffusion process. The correlation between the motion and control context is depicted by the cross-modal transformer encoder. Future motions are predicted autoregressively.

The first KL-divergence term with $q(y_S|y_0)||p(y_S)$ has no learnable parameters and can thus be ignored. $q(y_{s-1}|y_s, y_0)$ is conditioning on y_0 , which is tractable and can be represented as:

$$q(y_{s-1}|y_s, y_0) = \mathcal{N}(y_{s-1}; \tilde{\mu}_s(y_s, y_0), \tilde{\beta}_s \mathbf{I}), \quad (6)$$

where $\tilde{\mu}_s$ and $\tilde{\beta}_s$ is calculated as:

$$\begin{aligned} \tilde{\mu}_s(y_s, y_0) &= \frac{\sqrt{\bar{\alpha}_{s-1}}\beta_s}{1 - \bar{\alpha}_s} y_0 + \frac{\sqrt{\bar{\alpha}_s}(1 - \bar{\alpha}_{s-1})}{1 - \bar{\alpha}_s} y_s \\ \tilde{\beta}_s &= \frac{1 - \bar{\alpha}_{s-1}}{1 - \bar{\alpha}_s} \beta_s \mathbf{I}. \end{aligned} \quad (7)$$

Ho et al. [16] show that the second KL-divergence term can be calculated as:

$$\mathbb{E}_q \left[\frac{1}{2\beta_s} \|\tilde{\mu}_s(y_s, y_0) - \mu_\theta(y_s, s, x, c)\|^2 \right] + \lambda, \quad (8)$$

where λ is a constant value that does not depend on θ . We choose the reparameterization method shown in [16] that

$$\mu_\theta(y_s, s, x, c) = \frac{1}{\sqrt{\bar{\alpha}_s}} \left(y_s - \frac{\beta_s}{\sqrt{1 - \bar{\alpha}_s}} \epsilon_\theta(y_s, s, x, c) \right), \quad (9)$$

and the objective function is simplified to

$$\mathbb{E}_{\epsilon, y_0, y_s} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_s} y_0 + \sqrt{1 - \bar{\alpha}_s} \epsilon, s, x, c)\|^2, \quad (10)$$

where ϵ_θ is a network that predicts Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and is trained with the MSE loss. The training is performed for all steps $s \in [1, S]$. The complete training procedure with the simplified objective function is displayed as Algorithm 1, with the diffusion dropout later introduced in Section III-D. After the network ϵ_θ is trained, given past poses x , control input c , and $y_S \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we can synthesize future motions y_0 with the reverse generation process $y_{s-1} \sim p_\theta(y_{s-1}|y_s)$:

$$y_{s-1} = \frac{1}{\sqrt{\bar{\alpha}_s}} \left(y_s - \frac{\beta_s}{\sqrt{1 - \bar{\alpha}_s}} \epsilon_\theta(y_s, s, x, c) \right) + \sqrt{\beta_s} \mathbf{z}, \quad (11)$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $s \in [2, S]$ and $\mathbf{z} = \mathbf{0}$ when $s = 1$. We follow the generation procedure in Algorithm 2 to predict the next sample $y_0^t = \{m_0^{t+1}, \dots, m_0^{t+T_p}\}$. Inspired by Li

et al. [25] and Guillermo et al. [1], our model outputs the next T_p poses given the past T_h poses to improve model performance. We then only update the next time step $t + 1$, pass it autoregressively to the transformer-based encoder together with the next control input, and repeat the same procedure until all desired motion frames are synthesized.

D. Diffusion Data Dropout

Diffusion data dropout is applied during the training stage to improve data efficiency and model robustness. To be specific, we drop motion context information through the forward diffusion process without extra effort. A larger diffusion step s in the diffusion data dropout leads to the past motion x^t being more corrupted and less information is preserved. Similar to Equation 2, the past motion information with respect to step s can be calculated in one step as:

$$q(x_s^t|x^t) := \mathcal{N}(x_s^t; \sqrt{\bar{\alpha}_s} x^t, (1 - \bar{\alpha}_s) \mathbf{I}). \quad (12)$$

As shown in Algorithm 1, data dropout is done if $p < p_d$, where $p \sim \mathcal{U}(0, 1)$ and p_d is the diffusion dropout rate. We stabilize the training by starting with $p_d = 0$, i.e., with complete motion context information. A diffusion dropout scheduler $P_d = \{p_{d1}, p_{d2}, \dots, p_{dn}\}$ is set to increase the dropout rate during training. As the diffusion dropout rate increases, the denoising process becomes more focused on the control context and more robust to corruption of the motion context.

E. Motion Reconstruction

For motion reconstruction, we use the same framework without any further training. Our framework allows an imperfect input $\hat{x}^t = \{\hat{m}^{t-T_h}, \dots, \hat{m}^t\}$ since diffusion data dropout was applied during training. To reconstruct missing body joints or frames, we first generate a sequence of T_h future frames, $\{m^{t+1}, \dots, m^{t+T_h}\}$. Then we reverse the order of this sequence and the control signals $\{a^{t-T_h}, \dots, a^{t+T_h}\}$ to $\{m^{t+T_h}, \dots, m^{t+1}\}$ and $\{a^{t+T_h}, \dots, a^{t-T_h}\}$, exploiting the fact that the training data are augmented by lateral mirroring and time-reversion. We regard the reversed motion context and control context as conditioning information to

generate poses. The missing information can be reconstructed by filling the holes with the generated parts. We repeat until all missing parts are reconstructed.

F. Network Architecture

The proposed MoDiff framework is composed of an encoder and a decoder. The encoder encodes the past motion context and control context, and the decoder models the Gaussian transitions in the Markov chain, as depicted in Fig. 2. For the encoder, we design two-layer transformers that encode the motion context x^t and control context c^t separately, with output embeddings of the same dimension. The motion context is augmented by diffusion dropout. In the transformers, a position embedding is introduced to emphasize the positional relation at different time steps. The outputs are then concatenated together into a six-layers cross-modal transformer. For the decoder, we design a three-layer Transformer decoder similar to [18] to model spatial and temporal dependencies. The inputs include the future motion y^t , the noise variable ϵ , a diffusion step index s , as well as the output embedding of the cross-model Transformer. In diffusion step s , the noised future motion y_s^t is concatenated with a diffusion step embedding. Finally, fully-connected layers downsample the output to the motion dimension. Like in TimeGrad [24], we can pass autoregressively to the network to repeat until the desired horizon although TimeGrad primarily relies on recurrent nets for propagating temporal information.

IV. EXPERIMENTS AND EVALUATIONS

In this section, We first describe the experimental setting, including the dataset, ablation settings, and implementation details. We then present the results on the generation and reconstruction of human locomotion samples. On the basis of these, we evaluate and discuss the performance of both the proposed frameworks and the baselines, followed by preliminary results on various tasks.

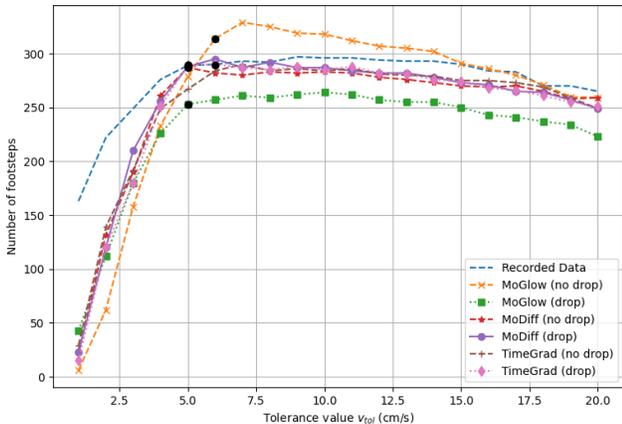


Fig. 3: Footstep analysis for generated motion given past information and control signals. The footstep count f_{est} on tolerance value v_{tol} . The black dots indicate v_{tol}^{95} , the first velocity tolerance for capturing 95% estimated footsteps.

A. Experimental Setting

Dataset. We evaluate MoDiff on a dataset of human locomotion preprocessed by [5] that retargets skeletons from the Edinburgh Locomotion MoCap, CMU Motion Capture, and HDM05 datasets, which includes various human gaits along different trajectories. The motion context $m \in \mathbb{R}^{63}$ is represented by the 3D coordinates of 21 body joints. The control context $a \in \mathbb{R}^3$ includes forward, lateral and rotational velocities. We slice the training data into 4-second clips and downsample these to 20 fps with 50% overlap. For synthesis and reconstruction with incomplete input, we generate data with missing parts by setting some joints to zero.

Ablation Settings. To assess the impact of design decisions, we compare our proposed MoDiff with MoGlow [5] and TimeGrad [24]. MoGlow is an autoregressive model based on normalizing flows and LSTM for human motion generation. TimeGrad is an autoregressive diffusion model for general time series forecasting and reports state-of-the-art performance on real-world datasets, such as traffic and electricity. Both MoDiff and MoGlow are probabilistic and designed for controllable motion synthesis. Unlike MoGlow, MoDiff and TimeGrad are diffusion-based models, but the encoder of TimeGrad is still based on LSTM, while MoDiff uses transformers. We evaluate the impact of the diffusion models and transformers architecture by comparing MoDiff with MoGlow and TimeGrad. To highlight the advantage of the proposed diffusion data dropout, we applied this dropout strategy to both MoDiff and the baseline models.

Implementation Details. Following the experimental setting in MoGlow [5], MoDiff is trained with 10-frame time windows, i.e., $T_h = 10$. In our experiments, the maximum number of diffusion steps S is 100. The diffusion dropout rate is $\{0, 0.05, \dots, 0.25\}$, increasing every 100 epochs after 500 epochs. The Transformer in our experiments is set to 256 dimensions and 4 heads. We use the standard Transformer encoder in PyTorch, with the T5-style relative positional embedding. We train with Adam optimizer with a learning rate of $5e-5$ and batch size of 64 for 1000 epochs. All experiments were conducted on a single NVIDIA A100 Tensor Core GPU.

B. Results and Discussions

Results. For quantitative analysis, we evaluate the quality of motion generation with footstep analysis and bone-length

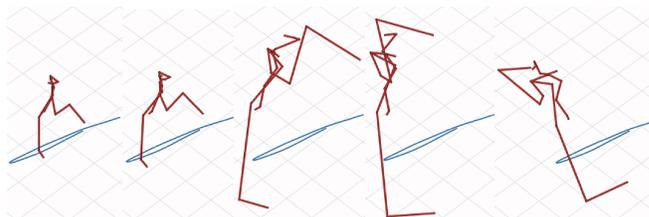


Fig. 4: Example poses with body joints flying away that are generated by MoGlow with incomplete past poses.

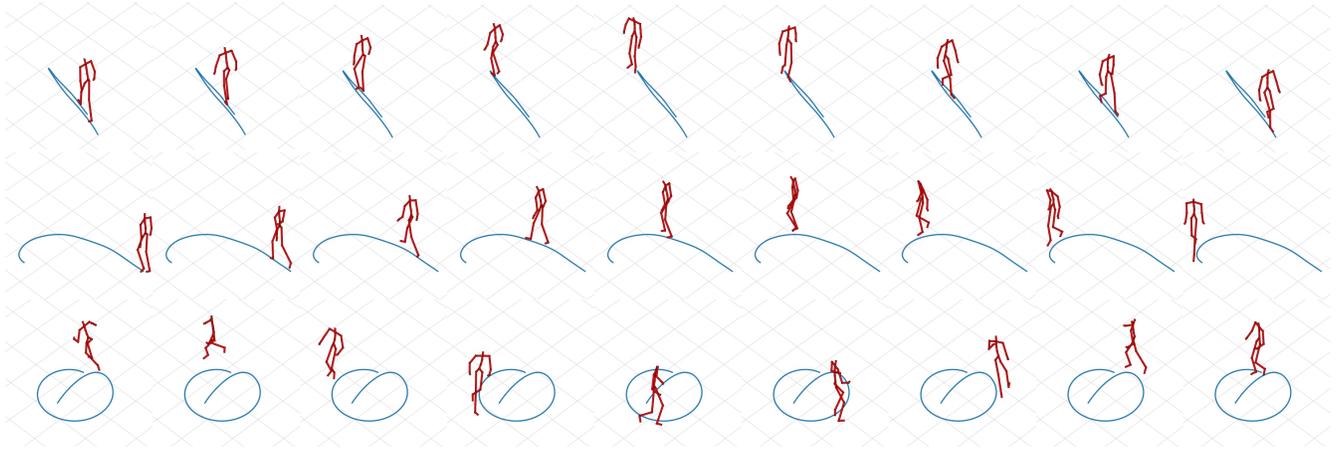


Fig. 5: Example sequences generated by MoDiff given different past information and control signals.

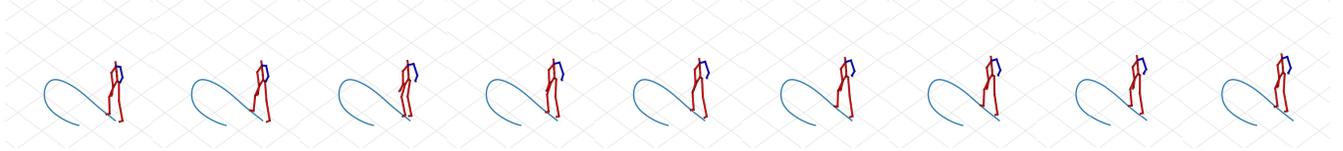


Fig. 6: An example sequence reconstructed by MoDiff. The right arm is missing in the past frames.

TABLE I: Results of the footsteps and bone-length analysis with complete input. The number closest to the recorded motion capture is in bold.

Model	f_{est}	v_{tol}^{95}	μ	σ	RMSE
Recorded Data	289	5	0.315	0.263	0
MoGlow (no drop)	314	6	0.222	0.134	0.315
MoGlow (drop)	253	5	0.341	0.250	0.250
TimeGrad (no drop)	284	6	0.331	0.338	0.124
TimeGrad (drop)	281	5	0.289	0.241	0.115
MoDiff (no drop)	280	5	0.289	0.239	0.089
MoDiff (drop)	284	5	0.334	0.286	0.072

TABLE II: Footsteps and bone-length results of generation and reconstruction with incomplete input. The number closest to the recorded motion capture is in bold.

Model	f_{est}	v_{tol}^{95}	μ	σ	RMSE _{ge}	RMSE _{re}
Recorded Data	289	5	0.315	0.263	-	-
MoDiff (no drop)	299	5	0.375	0.347	0.105	0.116
MoDiff (drop)	293	5	0.358	0.287	0.101	0.108

analysis, which are widely-used methods to evaluate artifacts related to heel sliding and limb stretching [5]. Footstep analysis compares the estimated footsteps of the generated motion and the recorded motion. The average number of footsteps f_{est} can be detected at intervals where the horizontal velocity of the heel joints is below a tolerance value v_{tol} . For evaluation, we compare the estimated number of footsteps at the first velocity tolerance for capturing 95% steps, denoted as v_{tol}^{95} . The estimated number of footsteps on different v_{tol} in the generated motions are displayed in Fig. 3, the detected v_{tol}^{95} is shown as black dot in this figure. We can observe that the curves of MoDiff and TimeGrad,

with and without diffusion dropout, are consistently closer to the curve of the recorded motion compared to MoGlow. Table I shows the results given complete context, include the estimated footsteps f_{est} , the tolerance value v_{tol}^{95} , the mean μ and standard deviation σ of the step duration, as well as the results of bone-length analysis. The bone-length analysis is for evaluating bone-stretching artifacts. We compare the RMSE (cm) of the bone length. Table II shows the results of generation and reconstruction given incomplete context.

Discussion. MoDiff can qualitatively and quantitatively generate recorded realistic motion with good diversity on the locomotion task, as demonstrated by the snapshots of generated motion sequences in Fig. 5. In contrast, the native MoGlow suffers from poor adherence and generates motion with noticeable foot-sliding artifacts. As can be seen from the results in Table I, the diffusion-based models, TimeGrad and MoDiff, perform significantly better in terms of both footstep and bone-length analysis compared to MoGlow which is based on normalizing flow. However, unlike the models the use LSTM, i.e. MoGlow and TimeGrad, MoDiff benefits from transformer-based encoders to further improve on bone-length analysis. The attention mechanisms in MoDiff discern specific features over a longer period, which is essential for learning consistent gaits under conditioning signals. Additionally, the attention-based transformer is effective at extracting meaningful information from intermediate diffusion steps with roughly corrected skeleton poses, which is not as feasible with standard LSTM models.

MoDiff can be extended to reconstruct incomplete body joints or frames without extra training, as illustrated by the reconstruction results in Table II, results that do not significantly diverge from those in Table I. When the past input is

incomplete, MoGlow exhibits unstable outputs, i.e. skeletons with joints flying away, as shown in Fig. 4. The results from MoDiff are closer to the recorded motion capture, illustrating the robustness of the diffusion-based architecture. From the example shown in Fig. 6, we observe that the reconstructed body joints (the right arm in blue) fit the original skeleton well. As can be seen from both Table I and Table II, the proposed data dropout strategy improves the performance of all three methods, which confirms its effectiveness.

Extra Applications. MoDiff is a task-agnostic framework that can be applied to other cross-modal generation tasks, such as text-to-motion and music-to-dance. We show generated samples in the attached multimedia materials link on YouTube: youtu.be/qeAs9eF3pbs. Note that the framework was not intended to be text-based even if it is agnostic to the input modality. The text here is just to generate a command token that can be dynamic and from other control modalities, e.g., keyboard strokes. It is thus more appropriate to compare MoDiff with other works for controllable locomotion generation.

V. CONCLUSIONS

We propose MoDiff, a Transformer-based diffusion model, to tackle the challenge of controllable and robust human motion synthesis and reconstruction under imperfect conditions. We introduce a novel diffusion data dropout strategy utilizing the diffusion forward process, which improves data efficiency and model robustness. The comparison results on the locomotion dataset with state-of-the-art baselines demonstrate the superiority of our MoDiff. MoDiff can be applied to various multimodal synthesis tasks. In the future, we plan to extend the MoDiff framework for classifier-guided conditional generation and apply it for more challenging dance motions.

ACKNOWLEDGMENT

This study has received funding from the European Commission Horizon 2020 research and innovation program under grant agreement number 824160 (EnTimeMent). This work benefited from access to the HPC resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

REFERENCES

- [1] G. Valle-Pérez, G. E. Henter, J. Beskow, A. Holzapfel, P.-Y. Oudeyer, and S. Alexanderson, "Transflower: probabilistic autoregressive dance generation with multimodal attention," *ACM TOG*, vol. 40, no. 6, pp. 1–14, 2021.
- [2] W. Yin, H. Yin, D. Kragic, and M. Björkman, "Graph-based normalizing flow for human motion generation and reconstruction," in *IEEE RO-MAN*. IEEE, 2021, pp. 641–648.
- [3] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *IEEE ICCV*, 2015, pp. 4346–4354.
- [4] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *IEEE CVPR*, 2017, pp. 6158–6166.
- [5] G. E. Henter, S. Alexanderson, and J. Beskow, "Moglow: Probabilistic and controllable motion synthesis using normalising flows," *ACM TOG*, vol. 39, no. 6, pp. 1–14, 2020.
- [6] Z. Wang, J. Chai, and S. Xia, "Combining recurrent neural networks and adversarial training for human motion synthesis and control," *IEEE TVCG*, vol. 27, no. 1, pp. 14–28, 2019.
- [7] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *arXiv*, 2022.
- [8] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv*, 2022.
- [9] Z. Li, Y. Zhou, S. Xiao, C. He, Z. Huang, and H. Li, "Auto-conditioned recurrent networks for extended complex human motion synthesis," *arXiv*, 2017.
- [10] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [11] L. Crnkovic-Friis and L. Crnkovic-Friis, "Generative choreography using deep learning," *arXiv*, 2016.
- [12] S. Levine, J. M. Wang, A. Haraux, Z. Popović, and V. Koltun, "Continuous character control with low-dimensional embeddings," *ACM TOG*, vol. 31, no. 4, pp. 1–10, 2012.
- [13] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura, "A recurrent variational autoencoder for human motion synthesis," in *BMVC*, 2017.
- [14] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose from speech with a conditional variational autoencoder." ISCA, 2017.
- [15] W. Yin, H. Yin, K. Baraka, D. Kragic, and M. Björkman, "Dance style transfer with cross-modal transformer," *arXiv*, 2022.
- [16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [17] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.
- [18] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu, "Stochastic trajectory prediction via motion indeterminacy diffusion," in *CVPR*, 2022, pp. 17 113–17 122.
- [19] G. Barquero, S. Escalera, and C. Palmero, "Belfusion: Latent diffusion for behavior-driven human motion prediction," *arXiv preprint arXiv:2211.14304*, 2022.
- [20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.
- [21] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *arXiv preprint arXiv:2211.09707*, 2022.
- [22] M. Zhang, C. Liu, Y. Chen, Z. Lei, and M. Wang, "Music-to-dance generation with multiple conformer," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 34–38.
- [23] J. Tseng, R. Castellon, and C. K. Liu, "Edge: Editable dance generation from music," *arXiv preprint arXiv:2211.10658*, 2022.
- [24] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *ICML*. PMLR, 2021, pp. 8857–8868.
- [25] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *ICCV*, 2021, pp. 13 401–13 412.
- [26] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv*, 2015.
- [27] X. Wang, S. Takaki, and J. Yamagishi, "Autoregressive neural f0 model for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1406–1419, 2018.
- [28] G. Kovács, L. Tóth, D. Van Compernelle, and S. Ganapathy, "Increasing the robustness of cnn acoustic models using autoregressive moving average spectrogram features and channel dropout," *Pattern Recognition Letters*, vol. 100, pp. 44–50, 2017.